SHORT COMMUNICATION

Communicative & Integrative Biology 5:6, 531–533; November/December 2012 © 2012 Landes Bioscience

Adapting without reinforcement

Aaron Kheifets* and C. Randy Gallistel

Department of Psychology; Rutgers University; Piscataway, NJ USA

Keywords: reinforcement learning, model-based control, probability estimation, timing, decision under uncertainty

Our data rule out a broad class of behavioral models in which behavioral change is guided by differential reinforcement. To demonstrate this, we showed that the number of reinforcers missed before the subject shifted its behavior was not sufficient to drive behavioral change. What's more, many subjects shifted their behavior to a more optimal strategy even when they had not yet missed a single reinforcer. Naturally, differential reinforcement cannot be said to drive a process that shifts to accommodate to new conditions so adeptly that it doesn't miss a single reinforcer: it would have no input on which to base this shift.

One of the oldest debates in all of learning is over how much of the environment subjects represent and how much of behavior is due to general-purpose algorithms such as hill-climbing. Classically, learning has thought to be gradual, trial-and-error process epitomized by the "learning curve" though numerous examples have been offered that suggest that this may be an artifact of averaging.[1] The processes supposed to underlie this classical view of learning are typically model-free, general-purpose methods in which the agent repeatedly tests various strategies for solving the problem they are confronted with.[2] The agent must then trade off between exploring new strategies, which may yield higher reward from their current one, and exploiting known good strategies. The particulars of this tradeoff and how to-be-tested strategies are selected is an enormous area of research both in psychology and AI.[3,4] In opposition is the view that agents model their environment and then optimize their behavior according to that model.[3,5] In Kheifets and Gallistel[6] we employed a task–already shown in Balci, Freestone and Gallistel[7] to produce near-optimal performance from subjects–that allowed us to inspect the time-course of learning, to see whether it followed a gradual evolution, as would be predicted for an algorithm that made small adjustments to its current strategy, or it made quantum leaps in its behavior, as would be predicted for an algorithm that models it's environment and calculates and implements a new optimal solution in response to a change in the environment.[8] We found strong evidence that not only is behavior model-based, it is not directly driven by an effort to maximize reward.

Ours was an interval-timing task in which subjects (mice of strain C57bl/6j) must use their estimate of the time elapsed from the beginning of the trial to decide whether to leave a short-latency feeding location (hopper) to go to a long-latency location (hopper) (see **Fig. 1**, taken from Kheifets and Gallistel[6]). The subject experiences two types of trials (long and short) but is not signaled which type the current trial is until it has ended (the trial type is determined by a draw from the Bernoulli distribution with probability p_s): on a short trial, the mouse gets a pellet for poking on the short-latency hopper after three seconds. If its first poke after three seconds is on the long-latency hopper, however, the trial ends without reinforcement. Similarly, on long trials the mouse is reinforced for making it's first poke after nine seconds have elapsed on the long-latency hopper but loses its reward if this poke is to the short side instead. Subjects get fed on nearly every trial by first going to the short side, waiting until they are certain that three seconds have elapsed and then moving to the long side.

Though switching hoppers at any time during the three-to-six second window would yield a reinforcer on that trial, the fact that subjects are not perfect timers introduces risk to attempting to switch very early or very late in the temporal window. Choosing the strategy to try to switch at 8.9 sec would result in a large percentage of trials in which the subject would accidentally switch too late (and similarly for attempting to switch at 3.1 sec). The task was designed such that the optimal target switch time was dependent on the proportion of probe trials to normal trials: when the proportion was high, the danger of switching too late was larger than the danger of switching too early (and vice-versa). We manipulated the percentage of long trials (p_s) and observed how subject shifted their switch behavior in response.

From simply looking at the scatter plot of our raw data, the time-course of learning made it clear that they strongly argued against classical, model-free theories of learning: subjects made abrupt shifts in their behavior and did so shortly after we manipulated the probability of a long or short trial. This basic result runs counter to models of learning that operate through small, incremental shifts in behavior, driven by the level of reinforcement the subject experiences when it makes these shifts (often supposed to be the result of Hebbian learning). Beyond this, we found that our data argued against reinforcement-driven models of learning altogether, be they model-based or model-free. One of the ways in which we demonstrated this was to use the fact that

*Correspondence to: Aaron Kheifets; Email: Kheifets@eden.rutgers.edu
Submitted: 06/13/12; Revised: 07/09/12; Accepted: 07/12/12
http://dx.doi.org/10.4161/cib.21474.

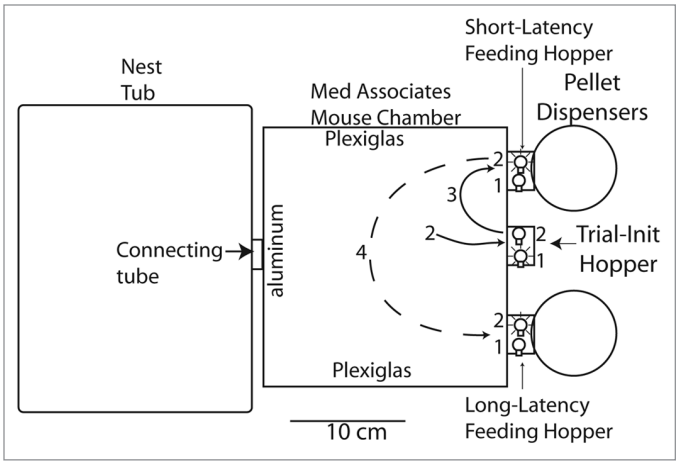

Figure 1. The experimental environment. In the Switch Task, a trial proceeds as follows: (1) Light in the Trial-Initiation Hopper signals that the mouse may initiate a trial. (2) Mouse approaches and pokes into the Trial-Initiation Hopper, extinguishing the light there and turning on the lights in the two feeding hoppers (trial onset). (3) Mouse goes to the short-latency hopper and pokes into it. (4) If, after 3 sec have elapsed since trial onset, poking in the short-latency hopper does not deliver a pellet, mouse switches to the long-latency hopper, where it gets a pellet there in response to the first poke at or after 9 sec since trial onset. Lights in both feeding hoppers extinguish either at pellet delivery or when an erroneously timed poke occurs. Short trials last about 3 sec and long trials about 9 sec, whether reinforced or not: If the mouse is poking in the short hopper at the end of a 3 sec trial, it gets a pellet and the trial ends. If it is poking in the 9 sec hopper, it does not get a pellet and the trial ends at 3 sec. Similarly, long trials end at 9 sec: If the mouse is poking in the 9 sec hopper, it gets a pellet; if in the 3 sec hopper, it does not. A switch latency is the latency of the last poke in the short hopper before the mouse switches to the long hopper. Only the switch latencies from long trials are analyzed. Reproduced from Figure 1 in Kheifets A, Gallistel CR. Mice take calculated risks. Proc Natl Acad Sci U S A 2012; 109:8776-9.

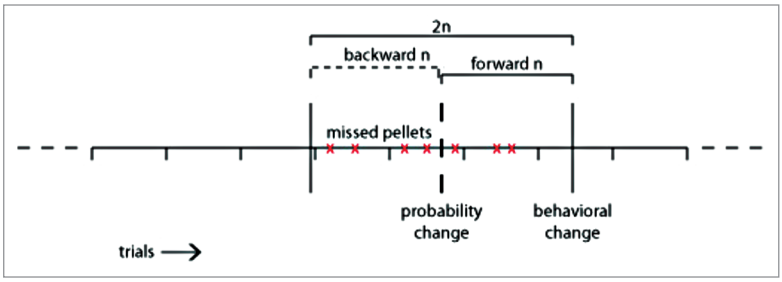

Figure 2. The number of trials to the midpoint of the behavioral shift is depicted by n here. One can then look at the n trials after the change in probability to find the number of reinforcers missed while the subject adjusts to the new probability of a long trial. We can then compare this number with the number of reinforcers missed in the n trials leading up to the change in probability (when the subject is already adjusted to the previous value) to see if that number is typically different from the number missed while the subject gathers evidence that a change has occurred and shifts its behavior accordingly. These two values are compared in **Figure 3**.

all reinforcement-learning models require a detectable difference in the level of reinforcement the subject receives before and after it shifts its behavior. For example, if my current strategy is to switch after 6 sec have elapsed and I wish to know whether or not I should increase that time, I would test out a slightly longer time and see whether this led to more or fewer reinforcements.

To test this, we looked at the number of reinforcers missed in the trials between when we changed the probability of a short trial and when the animal altered its behavior. It is only the experience of these trials that caused the behavioral shift. We then looked at the number of reinforcers missed in the same number of trials prior to when we changed the probability of a short trial (**Fig. 2**).

We found that the distributions for the pellets missed were indistinguishable (in **Fig. 3** you can see that they lie right on top of one another) and, perhaps most impressively, in a significant number of cases (here approximately 40%) the subjects shifted their behavior before they missed a single reinforcer. No theory that claims behavioral shifts are driven by slight changes in behavior garnering more reinforcements can explain this fact.

We do not argue that subject behavior was not based on maximizing reinforcement. On the contrary: it is because of the subjects' optimizing their earnings in the long run that we were able to definitively say that they are not shifting their behavior in response to reinforcement but rather in anticipation of the reinforcers based on their models' predictions. In this way, this is an example not only of model-based learning but also of learning that is not directly reinforcement-driven. One does not need to drive off the edge of the road several times to develop the behavior of driving comfortably away from the edge of the road.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. Gallistel CR, Fairhurst S, Balsam P. The learning curve: implications of a quantitative analysis. Proc Natl Acad Sci U S A 2004; 101:13124-31; PMID:15331782; http://dx.doi.org/10.1073/pnas.0404965101.
2. MacKay D. Information Theory, Inference, and Learning Algorithms. Cambridge, England: Cambridge Universiy Press, 2003.
3. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998.
4. Bishop C. Pattern Recognition and Machine Learning. New York: Springer, 2006.
5. Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. Cogn Affect Behav Neurosci 2008; 8:429-53; PMID:19033240; http://dx.doi.org/10.3758/CABN.8.4.429.
6. Kheifets A, Gallistel CR. Mice take calculated risks. Proc Natl Acad Sci U S A 2012; 109:8776-9; PMID:22592792; http://dx.doi.org/10.1073/pnas.1205131109.
7. Balci F, Freestone D, Gallistel CR. Risk assessment in man and mouse. Proc Natl Acad Sci U S A 2009; 106:2459-63; PMID:19188592; http://dx.doi.org/10.1073/pnas.0812709106.
8. Dayan P. Robust Neural Decision-Making. Evolving the mechanisms of decision making: Toward a darwinian decision theory. Frankfurt: Ernst Strüngmann Forum, 2012.

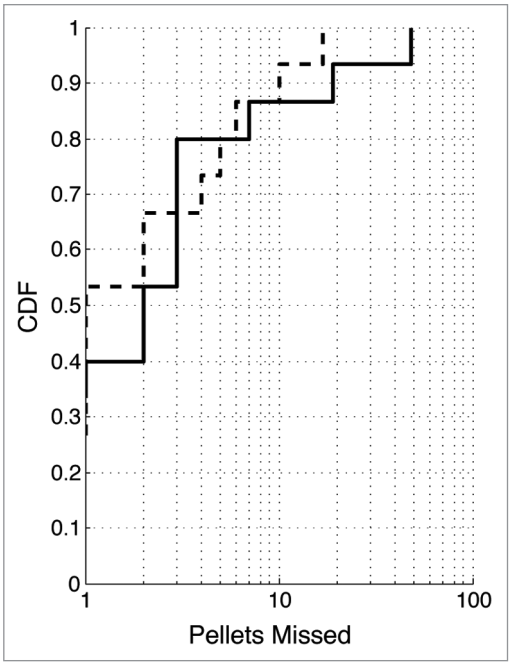

Figure 3. Cumulative distribution of the number of pellets missed after the probability value changed but before the animal was half way through its behavioral shift (solid line), and the same number of trials before the probability change (dashed line). Note that the two lie right on top of one another. Moreover, in the majority of instances, very few pellets were missed before the subject changed its behavior. In 40% of cases, not even a single pellet was missed before the subject was half way through its behavioral shift.